We would like to thank the reviewers for their generous comments and suggestions. We are encouraged that they found that our formalism is *extremely elegant* (R2) and *perfectly unifies the prior methods* (R3), and that our analyses were *sound and thorough* (R2, R3). Reviewer 4's assessment that **"this is potentially the definitive paper on the mathematical analysis of cooperative communication, establishing a firm foundation for this topic and providing a lot of ground on which to build."** is worthy of special note. We believe the strength of these endorsements support acceptance of the paper in NeurIPS, and encourage the reviewers to consider our responses in their decisions.

Summary. Recall that we present a unifying EOT framework for analyzing cooperative communication. Build upon machinery in optimal transport, we provide answers to fundamental questions about a broad **class of models** of cooperative communication. Specifically, we (a) theoretically guarantee the existence of optimal communication plan and algorithmically ensure the achievability of such plans by convergence of Sinkhorn scaling; (b) mathematically analyze and computationally verify the robustness to violations of common ground. These theoretical results are important because they establish viability in practice. It is implausible (even impossible) for any two agents to have exactly the common ground, if any model is to serve as a theory of human or machine behavior in realistic settings, besides the existence of optimal plans, the stability result is necessary. Moreover, our framework also provides strong links over a wide range of research topics—pragmatic reasoning, robotics, machine teaching, and Bayesian probability—that are important to the NeurIPS community.

(R4-Q1) "What is the relation between the current work and the literatures on cooperative inference [Yang et al., 2018, Wang et al., 2019]?" **Cooperative inference is a special case of our unified EOT framework** with greedy parameter $\lambda = 1$ (Prop. 3 in Sec.2.3, see detail derivation in Supplement Sec. B.1). The new, general formulation has several theoretical and practical implications. Theoretically, previously unproven results include: convergence of cooperative communication for *arbitrary* priors (Prop. 3), smoothness (in particular differentiability) of cooperative communication which implies the deviations in common ground are repairable (Prop.4& 5), analysis of instability under greedy data selection with difference choice of λ (Sec.3.3, Prop 6).

(R2) "[W]hether the proposed theory can have practical consequences or just a matter of unifying..." **Fundamentally, theoretical guarantees—of convergence, of algorithms, of robustness—have deeply practical implications**. For example, existing models assume, without basis, that models will work despite the fact that communicating partners can never have perfect common ground. Our analysis justifies practical application of the class of models, and suggests under when (in terms of parameterizations) and why violations may result in failures of communication.

(R4-Q2) "Maybe a table that spells out the formal correspondence between various existing cognitive models and components of EOT"? **Great suggestion!** We will include a such table in the revised supplemental materials.

(R4-Q3) "Connection with the information theoretic part in 2.4 needs to be better contextualized". **We will do so.** Cooperative communication, like standard information theory, involves communication over a channel. It is therefore interesting and important to ask whether there is a formal connection. Notably, the connection established in Sec.2.4 is an additional demonstration of the value of our general formulation.

(R3,R4-Q4) " [L]ack of experiments to really showcase the usefulness of this new unified model" &"[I]t would be nice to see the other results tied more closely to human behavior." **Our goal is establishing a mathematical framework for proving statements about the class of models**, rather than assessing a single model. To that end, our theoretical results are novel in both the machine learning and human learning literatures. Our simulation and model fitting results illustrate the implications of our analyses (Sec.3) by demonstrating predicted differences among specific models for both machine and human cooperative communication (Sec.4).

(R2, R3) "[I]gnore completely game-theoretic analysis of communication games" & "enable the agents to emerge some behaviors during cooperative communication, as in [1]." **We will acknowledge game theory**; however, we are unaware of game theoretic approaches that are competitive in both human and machine cooperative communication. **Our mathematical analysis is of single interactions, unlike POMDPs as in [1]**. Note that [1] has no proofs.

(R2) "[D]omains not of direct interest to machine learners". **We respectfully disagree, and we provide evidence.** The idea of cooperative communication has been proposed and applied in a wide range of existing machine learning models such as teaching by demonstration [Ho et al., 2016] and cooperative inverse reinforcement learning [Hadfield-Menell et al., 2016] which were both published in NeurIPS. In addition, NeurIPS has published papers on machine teaching (5 papers in NeurIPS 2019!), and machine-human or machine-machine collaboration and teaming. Our framework also bridges many topics central to the NeurIPS community (as measured by appearances in titles of papers): optimal transport (13 papers last year), Sinkhorn algorithm (3 papers) and Bayesian (40 papers) probability (14 papers).

We greatly appreciate the reviewers' enthusiastic endorsements, as well as their suggestions and comments. We will reorganize the contents and provide clear summary of the main justifications and advantages of the EOT framework to incorporate reviewers feedback.